ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Medical Records Clustering: A Survey

Mangesh Mali[1], Dr Parag Kulkarni[2], Prof. Virendra Bagade[3]

M.E. Student, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India[1]

Chief Scientist, Research Department, iknowlation Research Labs, Pune, India[2]

Asst. Professor, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India[3]

ABSTRACT: Retrieving similar medical cases from the medical case repository for user search case, the similarity measure and good clustering is useful. While To finding similarity between cases several methods have been proposed, but measuring the similarity between patient cases is a challenging problem. In that survey we focus on different similarity measures and clustering techniques. We are working on the data of medical records. Data is high dimensional, that much of features not gives much accuracy so we extract features from the medical records and build case library. We compare the result of different clustering algorithms using clustering validation.

KEYWORDS: Case-based reasoning, Extracting, User theme, Clustering, similarity measures, Clustering Validation

I. INTRODUCTION

With the advent of electronic health records, more data is continuously collected for individual patients, and more data is available for review from past patients. Despite this, it has not yet been possible to successfully use this data to systematically build computer-based decision support systems that can produce clinical recommendations to assist clinicians in providing individualized health-care. Medical Decision making should use relevant data from many distributed systems instead of a single data source to maximize its applicability but real-world medical data are often based on missing information. This is referred as the medical information challenge.

In past, the specialist applies their insight in the therapeutic choice and finding framework. After applying their insight they make a watchful treatment on the premise of patients clinical exam result in a blend of their history. There is the need to give precise determination and treatment to offer assistance in patient recuperation. Various variables which can impact customary restorative determination process are introduced. The data mining is broadly utilized as a part of PC based therapeutic analysis, which utilizes the medicinal cases to get the conclusion run the show.

Now days, large volume of data available in the medical system which gives the opportunity to construct computer based patient medical cases. Two issues are important in the construction of medical diagnosis decision system: the problem of constructing medical cases directly from raw data by imputing missing value and creating medical system with respect to user.

We focus on the creating system with respect to the user. Here we say that user could be the patient or doctor. The patient is concerned about symptoms, type of treatment and more, where the doctor is concerned about symptom study, possible causes related to new patient symptoms. In the proposed *system*, we analyze user (patient/doctor) search query and retrieve similar cases based on user theme. Case-based reasoning is the model which solves the problem by analyzing previously available cases and by reusing information and knowledge of the available cases. The System calculates the distance between search case and case in the cases repository using similarity measurement methods. Both case search and matching process need to be successful and time efficient.

The objective for this research is, to develop user interactive system to provide services to a user via an interactive search for personalized patient needs. To design appropriate structure for case content and indexing on case repository. Extract patient-specific features from medical cases which are the most describing case and useful for finding similarity metric. Apply similarity measures for retrieve cases relevant to search case and user theme from case repository. Recommend most similar cases to a user.

Some well know similarity measures are such as cosine similarity, Euclidean distance, Manhattan distance etc. These measures are used with different clustering algorithms such as DBSCAN, K-means, hierarchical clustering etc.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Wediscuss all clustering algorithms which are applicable to our datasets. After analyzing the results we take better algorithm for our research work.

II. RELATED WORK

In this paper [1] they present approach of combing abstracted patient-specific features medical cases. The information-theoretical measure is used to compute similarity between cases. It is efficient method to represent cases. They implemented two information-theoretical measure in this study are corpus-dependent weighing models (the Nats model and the Bin model). Both methods then tested by expert evaluation of case similarity. Some limitations of this paper first, study rely upon an abstracting system which abstracts the feature from medical text. Second, abstracted features are only applied into an informational-theoretic measure using two corpus-based weighing models.

In paper [2] bring into use a new distance measure that is that is better suited than tradition methods at detecting similarity in patient record by referring to a concept hierarchy. They measure the distance to new distance measure for categorical values by considering path distance between a concepts in hierarchy in an account. The new distance measure is an improvement over the current standard hierarchical arrangement of categorical values is available.

In the paper [3], the user probabilistic model to measuring the similarity between patient trace for clinical pathway analysis. Analysis of patient trace repository is unsupervised. Critical treatment behaviour can be discovered, analysed based on topic analysis given in that study.

In this paper [4] a Case-Based Reasoning Application was developed for treatment and management of diabetes using jCOLBIRI CBR framework. That application uses available past patient cases to present reasoning. The system employed case based methodology of reasoning which involves the process. The success of the system depends on the use of a similarity matching between the available cases and the new search case. The system deployed and tested with real life cases and then updated by a medical expert. The accuracy of the system can further be improved by combining different pattern matching algorithms such as (Euclidean, Hamming distance, neural networks etc.).

In paper [11], the author present a probabilistic approach to measuring the similarities between patients traces for client pathway analysis. They introduces three possible applications i.e., patient trace retrieval, clustering, and anomaly detection. To evaluate applications via real-world data-set of specific clinical data collected from a Chinese hospital. The patient traces could be measured based on their behavioural similarities.

In paper [12], the author designs a patient similarity framework which combines both unsupervised information and supervised information. They propose a novel patient similarity algorithm that uses spline regression to capture the unsupervised information. They also propose an algorithmic framework that could incrementally update the existing patient similarity measure from Patient similarity framework using matrix theory. They should speeds up the physician feedback and newly available clinical information by introducing a general on-line update algorithm for PSF matrix.

In paper [6], the author proposes a framework for the recommendation of the doctor and builds doctor profile. They firstly suggested for finding the similarities between user's consultation and doctor's profiles. Then, to measure doctor's quality, experiences, and different users opinions are considered. Finally, to combine the results of the relevance model and the quality model, and then recommended a doctor. A mobile recommender APP is proposed.

In these paper, [13] a survey is carried out to extract the new set of features efficiently. Here, many feature extraction algorithms proposed by different researchers are discussed and the issues present in the existing algorithm were identified. The future work of that study is to overcome the issues and to propose a new feature extraction algorithm, which will extract the new set of features and to improve the classification accuracy.

III. RECORD CLUSTERING

Cluster Analysis (data segmentation) has a variety of goals that relate to grouping or segmenting a collection of objects (i.e. observations, individuals, cases, or data rows) into subsets or clusters, such that those within each cluster are more closely related to one another than objects assigned to different clusters. Central to all of the goals of cluster analysis is the notion of degree of similarity (or dissimilarity) between the individual objects being clustered.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

K-means

K-means tries to improve the inter group similarity while keeping the groups as far as possible from each other. Basically K-Means runs on distance calculations, which again uses Euclidean Distance for this purpose. The basic restriction for K-Means algorithm is that your data should be continuous in nature. It won't work if data is categorical in nature. K-Means is an iterative process of clustering; which keeps iterating until it reaches the best solution or clusters in our problem space. But the basic question which should arrive is that how to decide the number of clusters (K). There is no mathematical formula which can directly give us answer to K but it is an iterative process where we need to run multiple iterations with various values of K.

Partitioning Around Medoids(PAM)

PAM is related to the *k-means* algorithm and the medoidshift algorithm. PAM is realisation of *k*-medoid clustering. PAM uses a greedy search which may not find the optimum solution. *k*-medoids algorithms are partitioning (breaking the dataset up into groups) and attempting to minimise the distance between points labelled to be in a cluster and a point designated as the centre of that cluster.

Hierarchical

In hierarchical clustering, the data is not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to n clusters that each contain a single object. Hierarchical Clustering is subdivided into agglomerative methods, which proceed by a series of fusions of the n objects into groups, and divisive methods, which separate n objects successively into finer groupings. Hierarchical clustering may be represented by a two-dimensional diagram known as a dendrogram, which illustrates the fusions or divisions made at each successive stage of analysis. Problems with hierarchical clustering- Computational complexity in time and space, Once a decision is made to combine two clusters, it cannot be undone, No objective function is directly minimized, Sensitivity to noise and outliers, Difficulty handling different sized clusters and convex shapes, Breaking large clusters.

DBSCAN

DBSCAN is a density-based algorithm uses density as number of points within a specified radius where point is a core point. This density-based algorithm eliminates noise points and makes each group of connected core points into a separate cluster. DBSCAN is resistant to noise and can handle clusters of various shapes and sizes. DBSCAN does not require one to specify the number of clusters in the data a priori, as opposed to k-means. DBSCAN can find arbitrarily shaped clusters. It can even find a cluster completely surrounded by (but not connected to) a different cluster. Due to the MinPts parameter, the so-called single-link effect (different clusters being connected by a thin line of points) is reduced. DBSCAN has a notion of noise. DBSCAN requires just two parameters and is mostly insensitive to the ordering of the points in the database. (However, points sitting on the edge of two different clusters might swap cluster membership if the ordering of the points is changed, and the cluster assignment is unique only up to isomorphism.)

IV. SIMILARITY MEASURES

A similarity measure can be defined as the distance between various record features. Similarity is amount that represents strength of relationship between them. Here, a brief overview of similarity measure function commonly used in clustering.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Cosine similarity

Cosine similarity computes the pairwise similarity between two documents using dot product and magnitude of vector document A and vector document B in high-dimensional space. The following formula calculates the Cosine similarity between vector (documents) A and vector B in n dimensional space:

$$V = cosine(A, B) = \frac{\sum_{n=1}^{k} A(n).B(n)}{|A|.|B|}$$

Where, A and B are the vector of document A and document B respectively in n dimensional space.

Euclidean distance

Euclidean distance (ED) is another geometrical measure used to measure similarity of two documents. Each document is represented as a point in space based on term frequency of n terms (representing n dimension). ED computes the difference between two points in n dimensional space based on their coordinate using following equation:

$$ED(A, B) = \sqrt{\sum_{n=1}^{k} (A(n) - B(n))^2}$$

Where, A and B are the vector of document A and document B respectively in n dimensional space.

Manhattan distance

Manhattan distance is a distance metric that calculates the absolute differences between coordinates of pair of data objects.

$$Dist_{XY} = |X_{ik} - Xjk|$$

Jaccard distance

The Jaccard distance measures the similarity of the two data items as the intersection divided by the union of the data items.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

We apply above clustering techniques on medical records. We get different clusters for different techniques. k-means algorithm gives the results but we have decide the value of k means number of clusters. And also there is no mathematical formula which can directly give us answer to K but it is an iterative process where we need to run multiple iterations with various values of K. Same as problem associated with PAM algorithm.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

DBSCAN is most popular clustering algorithm, while we apply on our dataset, it doesn't give not much effective results as expected. DBSCAN algorithm is the density based algorithm, in this algorithm objects get clustered depending on DBSCAN parameters which are epsilon and minPoints. So on different combinations of both parameters we don't get accurate clusters. Here the results are like- all objects are clustered only in one group or all objects are in its own cluster.

Hierarchical clustering gives better result than the above all. We applied clustering validation on these four algorithm results, we get graphs which shows better clustering algorithm for our dataset, see fig.1 and fig.2 for stability validation and internal validation respectively. Criteria column in Table 1 shows the value components work best on those values. E.g. In internal validation, component connectivity criteria is minimized, i.e. connectivity of hierarchical clustering is better than other algorithms, refer fig.2.

Validation	Components	Value	Criteria
Internal	Connectivity	[0, +∞]	Minimized
	Silhouette	[−1, +1]	Maximized
	Dunn Index	[0, +∞]	Maximized
Stability	APN (*average proportion non-overlap*)	[0, 1]	Minimized
	AD (*average distance*)	[0, +∞]	Minimized
	ADM (*average distance between means)*	[0, +∞]	Minimized

Table1. Summary of Validation Criteria using cluster validation

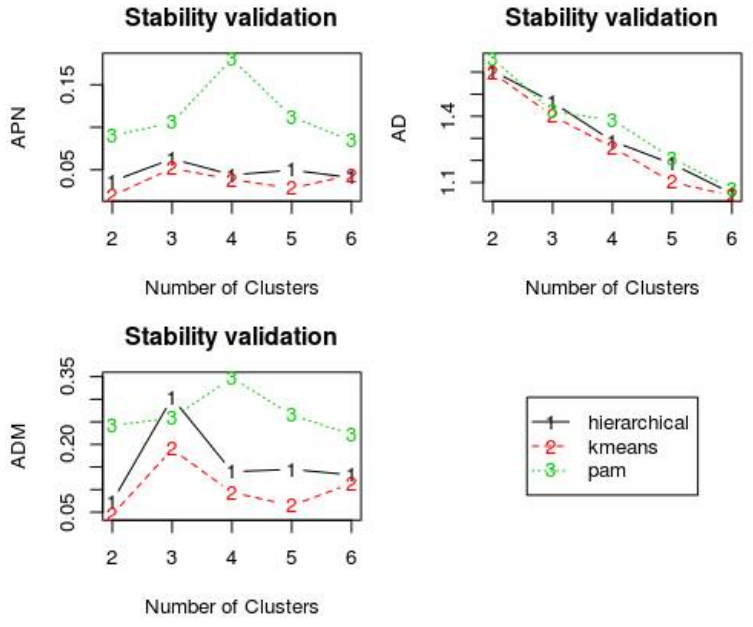

Fig.1. Stability validation

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

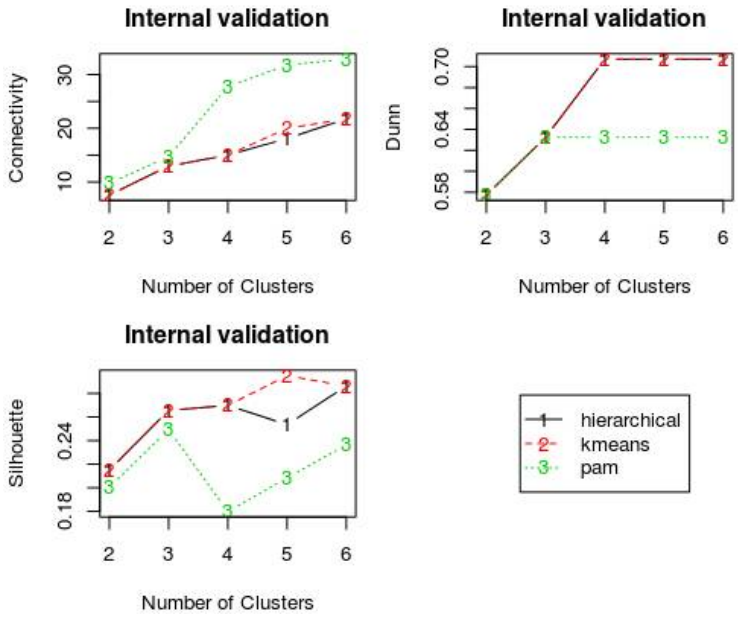

Fig.2. Internal validation

V. CONCLUSION

To conclude, this investigation found that, except for DBSCAN, the other clustering algorithms comparable effectiveness for our medical records dataset. From the observations we say that k-means gives the good clustering of objects comparable to other algorithms. Despite all above differences, we doesn't get accurate cluster if we check manually. So we are trying to introduce different clustering algorithm that gives accurate clusters on our data. Cosine similarity measure gives better performance than other measure.

REFERENCES

1. Hui Cao, Genevieve B. Melton, MarianthiMarkatou, George Hripcsak, "Use abstracted patient-specific features to assist an information-theoretic measurement to assess similarity between medical cases", Journal of Biomedical Informatics, 2008.
2. Dominic Girardi, Sandra Wartner, Gerhard Halmerbauer, Margit Ehrenmller, Hilda Kosorus, Stephan Dreiseitl, "Using concept hierarchies to improve calculation of patient similarity", Journal of Biomedical Informatics, 2016.
3. Zhengxing Huang, Wei Dong, HuilongDuan, HaominLi,"Similarity Measure Between Patient Traces for Clinical Pathway Analysis: Problem,Method, and Applications", IEEE Journal of Biomedical Informatics, VOL. 18, NO. 1, january 2014.
4. Mark K. Kiragu, Peter W. Waiganjo,"Case based Reasoning for Treatment and Management of Diabetes",International Journal of Computer Applications Volume 145- No.4, July 2016.
5. Maria Daltayanni, Chunye Wang, Ram Akella, "A Fast Interactive Search System for Healthcare Services", Service Research and Innovation Institute Global Conference, 2012.
6. Hongxun Jiang, Wei Xu, "How to find your appropriate doctor :An integrated recommendation framework in big data context", 2014.
7. SwarupanandaBissoyi, Brojo Kishore Mishra, ManasRanjanPatra, "ecommender Systems in a Patient centric Social Network - A Survey", International conference on Signal Processing, Communication, Power and Embedded System, 2016.
8. NabanitaChoudhury, ShahinAra Begum, "A Survey on Case-based Reasoning in Medicine",International Journal of Advanced Computer Science and Applications,Vol. 7, No. 8, 2016.
9. Jimeng Sun, Fei Wang, Jianying Hu, ShahramEdabollahi, "Supervised Patient Similarity Measure of Heterogeneous Patient Records", Volume 14, Issue 1, 2013
10. TaxiarchisBotsis, John Scott, Emily Jane Woo, Robert Ball, "Identifying Similar Cases in Document Networks Using Cross-Reference Structures", IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 19, NO. 6, NOVEMBER 2015.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

11. Zhengxing Huang, Wei Dong, HuilongDuan, Haomin Li, "Similarity Measure Between Patient Traces for Clinical Pathway Analysis: Problem, Method, and Applications",IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 18, NO. 1, JANUARY 2014.
12. Fei Wang, Jimeng Sun, "PSF: A Unified Patient Similarity Evaluation Framework Through Metric Learning With Weak Supervision", IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 19, NO. 3, MAY 2015.
13. N. Elavarasan, Dr. K. Mani, "A Survey on Feature Extraction Techniques",International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 1, January 2015.

BIOGRAPHY

Mangesh Mali is a student pursuing M.E. in the Computer Engineering Department, Pune Institute of Computer Technology, pune. His research interests are Data Mining, Data Analysis and Machine Learning.

Dr Parag Kulkarni is Chief Scientist and CEO of the iKnowlation Research Labs Pvt Ltd**,** an innovation, strategy and business consulting and product development organization. He has been visiting professor/researcher at technical and B-schools of repute including IIM, Masaryk University – Brno, COEP Pune.

Prof. Virendra Bagade is an Assistant Professor in the Computer Engineering Department, Pune Institute of Computer Technology, pune. His research interests are Data Mining and Data Warehouse, Information retrieval.